

# 2014-11-16 周报

杨哲

# 本周工作

## 数据处理

本周继续处理了上周关于微博的信息，能够爬取相应用户发表的微博状态。能够匹配出一定数量的用户。但是缺点在于，由于新浪微博的管制，爬取的速度非常慢，大概是5s爬取一个页面的速度。而且长时间爬取会被封号。另外由于使用的方法的限制，只能匹配出一小部分的微博账号。

# 本周工作

## 关于品牌的统计

品牌的统计比刚开始想的难度要大。主要的问题在于数据的不规整。开始的想法是想通过IMEI的比对直接找到相应的手机品牌，但是在网上找到的IMEI比对数据都不全，不能做到有效的比对。然后的想法是通过分析user-agent来寻找手机品牌。首先是以user-agent相同的作为同一型号的手机，统计了user-agent的情况。但是发现user-agent的数据非常不规整，30多个品牌的手机user-agent的类型有460000多种，然后根据手机品牌的关键字查找合并user-agent，仍有80000条记录未能合并。然后爬取了中关村手机的型号和品牌对应表，但是数据依然是不规整，如果想处理为规整的数据，需要很长时间的人工参与。

相应的文件放在周报下面的文本里。

# 下周工作

下周由于要换数据，所以这些工作就到这里了。下周会看一下关于在数据集成部分存在不一致和模糊性问题的解决方案。